University of Hyogo at NTCIR-11
TaskMine by Dependency Parsing

Takayuki Yumoto
University of Hyogo, Japan
input: 花粉症対策をする
(ease hay fever)

Web

quadruplets
• (マスク(mask), を, つける(wear), -)

postpositional particle
negation

output: subtasks
• マスクをつける (wear a mask)
• 目薬を使う(use a eyewash)
• ...
Our Approach

1. Collecting pages
   – collecting seed pages by query modification
   – collecting detailed pages by link anchor

2. Extracting quadruplets
   – Finding operation (and negation)
   – Finding target and postpositional particle

3. Ranking quadruplets
   – Synonyms by Wikipedia corpus
   – ranking by site frequency
1. Collecting Pages

1. collecting seed pages by query expansion

- query string
  - "方法(method)"
- e.g. ease hay fever method

2. collecting detailing pages

- detailing pages
  - same domain
  - all query keywords in anchor text (and a title of the seed page)
2.1 Finding operation (and negation)

1. The end of sentence is the first candidate.

2. The declinable chunk depending on the other candidates is also a candidate.

Negation is extracted from a candidate chunk if it exists.
2.2 Finding target and postposition

the indeclinable chunk depending to operation → target and postpositional particle are extracted

extracted quadruplet

(マスク(mask), を, つける(wear), -)

(no) negation
3. Ranking by site frequency

• We use site frequency instead of DF to reduce effects of site template (e.g. copyright statement)

• We propose two ranking methods:
(A) Site frequency of pair
(マスク(mask), を, つける(wear), -)
→ Order by SF(pair of target and operation)
(B) Site frequency of min and max of target and op consider importance of target and op separately
(マスク(mask), を, つける(wear), -)
→ Order by Min(SF(target), SF(op)), Max(SF(target), SF(op))
To identify synonyms, we used Wikipedia data. If two words are used as anchor text linking to the same page, they are regarded as synonyms.
We tried 50 queries of 4 categories
We compared nDCG@k

<table>
<thead>
<tr>
<th>Category</th>
<th>k=1</th>
<th>k=5</th>
<th>k=10</th>
<th>k=50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>0.140</td>
<td>0.141</td>
<td>0.173</td>
<td>0.191</td>
</tr>
<tr>
<td>Education</td>
<td>0.000</td>
<td>0.075</td>
<td>0.107</td>
<td>0.161</td>
</tr>
<tr>
<td>Daily life</td>
<td>0.167</td>
<td>0.153</td>
<td>0.150</td>
<td>0.154</td>
</tr>
<tr>
<td>Sequential</td>
<td>0.100</td>
<td>0.237</td>
<td>0.253</td>
<td>0.259</td>
</tr>
</tbody>
</table>
Problems

• the part where subtasks are extracted 
  *e.g.* the page describing not only *methods* to ease hay fever but also *mechanism of hay fever*

• same subtask in different expressions 
  *e.g.* “wear a mask” = “use a mask”

• limitation of model : multiple targets are sometimes needed in a single subtask 
  *e.g.* 種に傷をつける (scratch a seed)
Our method consists of:

• Collecting pages by query modification and link anchor

• Extracting quadruplets by dependency parsing
  – Finding operation (and negation)
  – Finding target and postpositional particle

• Ranking by site frequency
  – Synonyms by Wikipedia corpus

Our approach is better than the baseline, but it should be improved